Predicting What You Already Know Helps: Provable Self-Supervised Learning

Jason D. Lee, Qi Lei, Nikunj Saunshi and Jiacheng Zhuo

March 2, 2021

Outline

Motivation

- Self-supervised representation learning solves auxiliary prediction tasks (known as pretext tasks), that do not require labeled data, to learn semantic representations. These pretext tasks are created solely using the input features, such as predicting a missing image patch, recovering the color channels of an image from context.

- Predicting this known information helps in learning representations effective for downstream prediction tasks.

Question

- What conceptual connection between pretext and downstream tasks ensures good representations?
- What is a good way to quantify this?

Example

Notation

X_1: input sample, X_2: pretext tasks, Y : downstream label.

Image Colorization:

- X_1: input image
- X_2: the background color of the images
- Y : class name - desert, sky, or forest.

Given knowledge of the label Y , one can possibly predict the background X_2 without knowing much about X_1. In other words, X_2 is approximately independent of X_1 conditional on the label Y.

Intuition

In the above approximate conditional independence settings, the only way to solve the pretext task is to first implicitly predict Y and then predict X_2 from Y. Even though there is no labeled data, the information of Y is hidden in the prediction for X_2.

Contribution of the paper

Proposing a mechanism based on conditional independence (CI) to explain why solving pretext tasks created from known information can learn representations useful for downstream tasks. We theoretically demonstrate the reduced downstream sample complexity achieved by self-supervised learning under this assumption.

Notations

- x: scalar quantities, $\boldsymbol{x}$: vector values, X: random variables, $\boldsymbol{X}$: matrices.
- We use standard $\mathcal{O}$ notation to hide universal factors, and $\tilde{\mathcal{O}}$ to hide log factors.
- $\|\cdot\|$ stands for l_2-norm or Frobenius norm for vectors and matrices.
- $E^L[Y|X]$ denotes the best linear predictor of Y given X.

$$E^L[Y|X=\boldsymbol{x}] := \boldsymbol{W}^\star \boldsymbol{x} + \boldsymbol{b}^\star, \quad \text{where} \quad \boldsymbol{W}^\star, \boldsymbol{b}^\star := \operatorname{argmin}_{\boldsymbol{W},\boldsymbol{b}} \mathbb{E}[\|Y - \boldsymbol{W}X - \boldsymbol{b}\|^2].$$

- Covariance matrix. For random variables X, Y , we denote Σ_{XY} to be covariance matrix of X and Y. The partial covariance matrix between X and Y given Z is:

$$\Sigma_{XY|Z} := cov\{X - E^L[X|Z], Y - E^L[Y|Z]\} = \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$$

Outline

Setup and Methodology

Let $X_1 \in \mathcal{X}_1 \subset \mathbb{R}^{d_1}$, $X_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_2}$ and $Y \in \mathcal{Y} \subset \mathbb{R}^k$. If Y is finite with $|Y| = k$, we assume $Y \subset \mathbb{R}^k$ is the one-hot encoding of the labels. $P_{X_1 X_2 Y}$ denotes the joint distribution over $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$. $P_{X_1 Y}$, P_{X_1} denote the corresponding marginal distributions.

The proposed self-supervised learning procedure is as follows:

Step 1 ((pretext task))
Learn representation $\psi(x_1)$ through $\psi := \min_{f \in \mathcal{H}} \mathbb{E}\|X_2 - f(X_1)\|^2$, where $\mathcal{H}$ can be different choices of function classes.
Step 2 ((downstream task)) Perform linear regression on Y with $\psi(X_1)$, i.e., $g(x_1) := \boldsymbol{W}^{\star\top}\psi(x_1)$, where $\boldsymbol{W}^\star = \operatorname{argmin}_{\boldsymbol{W}} \mathbb{E}_{X_1,Y}[\|Y - \boldsymbol{W}^\top \psi(X_1)\|^2]$. Namely, we learn $g(\cdot) = \mathbb{E}^L[Y|\psi(\cdot)]$.

Approximation Error & Estimation Error

Approximation Error: It measures how well ψ can do with when given infinite samples for the task.
Denote $e_{apx}(\psi) = \min_{\boldsymbol{W}} E[\|f^\star(X1) - \boldsymbol{W}\psi(X1)\|^2]$ with $f^\star = E[Y|X_1]$ is the optimal predictor for the task.
Estimation Error: It measures the sample complexity of ψ on the downstream task.
We assume access to n_2 i.i.d. samples $(\boldsymbol{x}_1^{(1)}, y^{(1)}), \cdots, (\boldsymbol{x}_1^{(n_2)}, y^{(n_2)})$ drawn from joint distribution with density $P_{X_1 Y}$. Write
$\psi(\boldsymbol{X}^{down}) = [\psi(\boldsymbol{x}_1^{(1)})|\psi(\boldsymbol{x}_1(2))|\cdots|\psi(\boldsymbol{x}_1^{(n_2)})]^\top \in \mathbb{R}^{n_2 \times d_2}$. that is applied row-wise on each sample. Given these samples, we do linear regression on top of the learned representation ψ and are interested in the excess risk that measures generalization.

$$\hat{\boldsymbol{W}} = \text{argmin}_{\boldsymbol{W}}\{\frac{2}{n}\|\boldsymbol{Y} - \psi(\boldsymbol{X}^{down})\boldsymbol{W}\|^2\}; \quad ER_\psi(\hat{W}) := E\|f^\star(X_1) - \hat{\boldsymbol{W}}^\top\psi(X_1)\|^2.$$

Outline

Recipe and Data assumption

In this section, we focus on the case when input X_1 and pretext target X_2 are conditional independence (CI). The general recipe for the results will follow the following steps:

- Find a closed-form expression for the optimal solution $\psi^\star$ for the pretext task.
- Use conditional independence to argue that $e_{apx}(\psi^\star)$ is small.
- Exploit the low rank structure of $\psi^\star$ to get a good sample complexity on downstream tasks.

Data assumption:

$$Y = f^\star(X_1) + N,$$

where $f^\star = \mathbb{E}[Y|X_1]$; We assume N is σ^2-subgaussian. For simplicity, we assume non-degeneracy in random variables: $\Sigma_{X_iX_i}$ and Σ_{YY} are full rank.

Assumptions and optimal linear predictor

Assumption (Jointly Gaussian)

X_1, X_2, Y *are jointly Gaussian*

Assumption (Conditional independence)

$X_1 \perp X_2 | Y$.

Under the above assumptions, we have

$$f^\star(\boldsymbol{x}_1) = \mathbb{E}^L[Y|X_1 = \boldsymbol{x}_1] = \Sigma_{YX_1}\Sigma_{X_1X_1}^{-1}\boldsymbol{x}_1$$
$$\psi^\star(\boldsymbol{x}_1) = \mathbb{E}^L[X_2|X_1 = \boldsymbol{x}_1] = \Sigma_{X_2X_1}\Sigma_{X_1X_1}^{-1}\boldsymbol{x}_1$$

Approximation error

Lemma (Approximation error)

Under Assumptions A.1 and A.2, if $\Sigma_{X_2 Y}$ has rank k, then we have $e_{apx}(\psi^\star) = 0$.

Note that $cov(X_1, X_2|Y) = \Sigma_{X_1 X_2} - \Sigma_{X_1 Y}\Sigma_{YY}\Sigma_{YX_2} = 0$, we have

$$\psi(x_1) = \mathbb{E}^L[X_2|X_1 = x_1] = \Sigma_{X_2 X_1}\Sigma_{X_1 X_1}^{-1} x_1 = \Sigma_{X_2 Y}\Sigma_{YY}^{-1}\mathbb{E}^L[Y|X_1].$$

As $\Sigma_{X_2 Y}$ is rank k, left inverse of $\Sigma_{X_2 Y}$ exists. Hence,

$$\mathbb{E}^L[Y|X_1 = x_1] = \Sigma_{X_2 Y}^{\dagger}\Sigma_{YY}\psi(x_1).$$

Remark

$\Sigma_{X_2 Y}$ being full column rank infers that $\mathbb{E}[X_2|Y]$ is of rank k,i.e., X_2 depends on all directions of Y . This roughly means that X_2 captures all directions of information of Y . This is a necessary assumption for X_2 to be a reasonable pretext task for predicting Y .

Estimation error

Theorem (Estimation error)

Fix a failure probability $\delta \in (0,1)$. Under Assumptions A.1 and A.2,, if additionally $n2 >> k + log(1/\delta)$, the excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}\psi^\star(\boldsymbol{x}_1)$ on the target task satisfies

$$ER_{\psi^\star}(\hat{\boldsymbol{W}}) \leq \mathcal{O}\left(\frac{Tr(\Sigma_{YY|X_1})(k + \log(k/\delta))}{n_2}\right),$$

with probability at least $1 - \delta$.

Compared to directly using X_1 to predict Y, self-supervised learning reduces the sample complexity from $\tilde{\mathcal{O}}(d_1)$ to $\tilde{\mathcal{O}}(k)$.

Conditional Independent Given Latent Variables

Practically, we may only observe part of Y.

Assumption (Conditional Independent Given Latent Variables)

There exists some latent variable $Z \in \mathbb{R}^m$ such that $X_1 \perp X_2 | \bar{Y}$,and $\Sigma_{X_2 \bar{Y}}$ is of rank $k+m$, where $\bar{Y} = [Y, Z]$.

Corollary

Under Assumption 1 and 3, the approximation error $e_{apx}(\psi^\star)$ is 0, Theorem 1 can be generalized by replacing k by $k+m$.

Conditional independence

Assumption

Let $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$ be random variables from some unknown distribution. Let label $Y \in \mathcal{Y}$ be a discrete random variable with $k = |Y| < d_2$. We assume conditional independence: $X_1 \perp X_2 | Y$.

Universal Function Class

Suppose we learn ψ from a function class $\mathcal{H}$ with universal approximation power, the optimal function is

$$\psi^\star(\boldsymbol{x}_1) = \mathbb{E}[X_2|X_1 = \boldsymbol{x}_1].$$

Lemma (approximation error)

Suppose random variables X_1, X_2, Y satisfy Assumption 4, and matrix $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y}\times d_2}$ with $\boldsymbol{A}_y := \mathbb{E}[X_2|Y = y]$ is of rank $k = |\mathcal{Y}|$. Then $e_{apx}(\psi^\star) = 0$.

Proof

Proof.

$$\begin{aligned}\psi(\cdot) := \mathbb{E}[X_2|X_1] &= \mathbb{E}[\mathbb{E}[X_2|X_1, Y]|X_1] \\ &= \mathbb{E}[\mathbb{E}[X_2|Y]|X_1] \\ &= \sum_y P(Y = y|X_1)\mathbb{E}[X_2|Y = y] \\ &= f(X_1)^\top \boldsymbol{A},\end{aligned}$$

where $f : \mathbb{R}^{d_1} \to \Delta_{\mathcal{Y}}$ satisfies $f(x_1)_y = P(Y = y|X_1 = x_1)$, and $\mathbf{A} \in R^{\mathcal{Y} \times d_2}$ satisfies $\mathbf{A}_y := \mathbb{E}[X2|Y = y]$. Here Δ_d denotes simplex of dimension d, which represents the discrete probability density over support of size d. Then,

$$\mathbb{E}[Y|X_1 = x_1] = \boldsymbol{Y} f(x_1) = (\boldsymbol{Y} A^\dagger)\psi(X_1).$$

Let $W^\star = \boldsymbol{Y} A^\dagger$ we complete the proof. □

Estimation Error

We make an assumption about the whitened data $\psi^\star(X_1)$ to ignore scaling factors. Note that all bounded random variables satisfy sub-gaussian property.

Assumption

We assume the whitened feature variable $U := \Sigma_\psi^{-1/2}\psi(X_1)$ *is a* ρ^2*-subgaussian random variable, where* $\Sigma_\psi = \mathbb{E}[\psi(X_1)\psi(X_1)^\top]$.

Theorem (General Conditional Independence)

Fix a failure probability $\delta \in (0,1)$*, under Assumption 4 and 5, if additionally* $n >> \rho^4(k + log(1/\delta))$*, then the excess risk of the learned predictor* $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}\psi^\star(\boldsymbol{x}_1)$ *on the downstream task satsifies:*

$$ER_{\psi^\star}(\hat{\boldsymbol{W}}) \leq \mathcal{O}(\frac{(k + \log(k/\delta))}{n_2}\sigma^2).$$

Function Class Induced by Feature Maps

Given feature map $\phi_1 : \mathcal{X}_1 \to \mathbb{R}^{D_1}$, we consider the function class $\mathcal{H} = \{\psi : \mathcal{X}_1 \to \mathbb{R}^{d_2} | \exists \boldsymbol{B} \in \mathbb{R}^{d_2 \times D_1}, \psi(\boldsymbol{x}_1) = \boldsymbol{B}\phi_1(\boldsymbol{x}_1)\}$. The optimal function in $\mathcal{H}$ is given by

$$\psi^\star(\boldsymbol{x}_1) = \Sigma_{X_2\phi_1}\Sigma^{-1}_{\phi_1\phi_1}\phi_1(\boldsymbol{x}_1).$$

Since $\psi^\star$ is a linear function of ϕ_1, it cannot have smaller approximation error than ϕ_1. However CI will ensure that $\psi^\star$ has the same approximation error as ϕ_1 and enjoys much better sample complexity.

Lemma (Approximation error)

If Assumption 4 is satisfied, and if the matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{Y} \times d_2}$ $\boldsymbol{A}_y := \mathbb{E}[X_2|Y = y]$ *is of rank* $k = |\mathcal{Y}|$. *Then* $e_{apx}(\psi^\star) = e_{apx}(\phi_1)$.

Function Class Induced by Feature Maps

CI with approximation error

Assumption (Bounded approximation error)

We assume

$$\|\Sigma_{\phi_1\phi_1}^{-1/2}\phi_1(X_1)a(X_1)^\top\|_F \leq b_0\sqrt{k}$$

almost surely.

Theorem (CI with approximation error)

Fix a failure probability $\delta \in (0,1)$, under Assumption 4, 5 and 6, if additionally $n_2 >> \rho^4(k + log(1/\delta))$, then the excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}\psi^\star(\boldsymbol{x}_1)$ on the downstream task satsifies:

$$ER_{\psi^\star}(\hat{\boldsymbol{W}}) \leq e_{apx}(\phi_1) + \mathcal{O}(\frac{(k + \log(k/\delta))}{n_2}\sigma^2).$$

Remark

- We note that since $X_1 \perp X_2 | Y$ ensures $X_1 \perp h(X_2) | Y$ for any deterministic function h, we could replace X_2 by $h(X_2)$ and all results hold. Therefore in practice, we could use $h(\psi(X_1))$ instead of $\psi(X_1)$ for downstream task.
- The last theorem is also true with Assumption 3 instead of exact CI, if we replace k by $k + m$. Therefore with self-supervised learning, the required labels are reduced from complexity for $\mathcal{H}$ to $\mathcal{O}(k)$ or $\mathcal{O}(k + m)$ depending on the condition.

Outline

Beyond CI

- In the previous slides, we focused on the case where exact CI is satisfied.
- We start with the jointly-Gaussian case, where approximate CI is quantified by partial covariance matrix.
- We then generalize the results and introduce covariance operator to measure approximate CI.

Warm-up Jointly Gaussian Variables

Assumption (Approximate Conditional Independent Given Latent Variables)

Assume there exists some latent variable $Z \in \mathbb{R}^m$ *such that*

$$\|\Sigma_{X_1X_1}^{-1/2}\Sigma_{X_1X_2|Y}\|_F \leq \epsilon,$$

$\sigma((\Sigma_{Y\bar{Y}}^{\dagger}\Sigma_{\bar{Y}X_2}) = \beta > 0$ *and* $\Sigma_{X_2\bar{Y}}$ *is of rank* $k+m$*, where* $\bar{Y} = [Y, Z]$.

Theorem

Under the above assumption with constant ϵ *and* β*, then the excess risk satisfies*

$$ER_{\psi^\star}(\hat{W}) \leq \frac{\epsilon^2}{\beta^2} + Tr(\Sigma_{YY|X_1})\frac{(d_2 + \log(d_2/\delta))}{n_2}).$$

Feature map and Reproducing Kernel Hilbert Space

Let $L_2(P_X)$ denotes the square integrable function with respect to measure P_X, we are interested in some function class $\mathcal{H}_X \subset L^2(P_X)$ that is induced by some feature maps:

Definition (General and Universal feature Map)

We denote feature map $\phi : \mathcal{X} \to \mathcal{F}$ that maps from a compact input space $\mathcal{X}$ to the feature space $\mathcal{F}$. $\mathcal{F}$ is a Hilbert space associated with inner product: $\langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$. The associated function class is: $\mathcal{H}_x = \{h : \mathcal{X} \to \mathbb{R} | \exists w \in \mathcal{F}, h(\boldsymbol{x}) = \langle w, \phi(\boldsymbol{x}) \rangle_{\mathcal{F}}, \forall \boldsymbol{x} \in \mathcal{X}\}$. We call ϕ universal if the induced $\mathcal{H}_x$ is dense in $L^2(P_X)$.

For $f, g \in \mathcal{H}_x$, we can find $w_f, w_g \in \mathcal{F}$ such that $f(\boldsymbol{x}) = \langle w_f, \phi(\boldsymbol{x}) \rangle_{\mathcal{F}}$ and $g(\boldsymbol{x}) = \langle w_g, \phi(\boldsymbol{x}) \rangle_{\mathcal{F}}$. We define

$$\langle f, g \rangle_{\mathcal{H}} = \langle w_f, w_g \rangle_{\mathcal{F}}.$$

Cross-Covariance operator:

Definition

(Cross-covariance operator). For random variables $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ with joint distribution $P : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and associated feature maps ϕ_x and ϕ_y, we denote by

$$C_{\phi_x \phi_y} = \mathbb{E}[\phi_x(X) \otimes \phi_y(Y)] = \int_{\mathcal{X} \times \mathcal{Y}} \phi_x(x) \otimes \phi_y(y) dP(x, y),$$

the (un-centered) cross-covariance operator. Similarly we denote by $C_{X\phi_y} = \mathbb{E}[X \otimes \phi_y(Y)] : \mathcal{F}_y \to \mathcal{X}$.

Understanding of Covariance operator:

Let X and Y are random variable on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Denote $\mathcal{H}_x$ and $\mathcal{H}_y$ be functional spaces induced by ϕ_x and ϕ_y. For $f \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$, we have the bilinear form

$$C(f, g) = cov[f(X), g(Y)].$$

By the Riesz representation theorem, C corresponds to an unique linear operator $\mathcal{C}_{XY} : \mathcal{H}_y \to \mathcal{H}_x$ by

$$\langle f, \mathcal{C}_{XY} g \rangle_{\mathcal{H}_x} = C(f, g).$$

Accordingly, we can define $\mathcal{C}_{\phi_x \phi_y}$ such that

$$\langle w_f, \mathcal{C}_{\phi_x \phi_y} w_g \rangle_{\mathcal{F}_x} = \langle f, \mathcal{C}_{XY} g \rangle_{\mathcal{H}_x}$$

Covariance operator, conditional independence and HS-norm

Lemma

With one-hot encoding map ϕ_y and arbitrary ϕ_1, $X_1 \perp X_2 | Y$ ensures:

$$\mathcal{C}_{\phi_1 x_2 | \phi_y} := \mathcal{C}_{\phi_1 x_2} - \mathcal{C}_{\phi_1 \phi_y} \mathcal{C}_{\phi_y \phi_y}^{-1} \mathcal{C}_{\phi_y x_2} = 0.$$

For covariance operator $\mathcal{C} : \mathcal{F}_y \to \mathcal{F}_x$, we define the HS-norm as

$$\|\mathcal{C}\|_{HS}^2 = \sum_{i,j} \langle \xi_i, \mathcal{C} \eta_j \rangle,$$

with ξ_i and η_j the complete orthonormal systems of $\mathcal{F}_x$ and $\mathcal{F}_y$.

Beyond CI (General Case)

Let $\boldsymbol{X}_1^{pre} = [\boldsymbol{x}_1^{(1,pre)}, \boldsymbol{x}_1^{(2,pre)}, ..., \boldsymbol{x}_1^{(n_1,pre)}]^\top \in \mathbb{R}^{n_1 \times d_1}$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_2^{(1)}, \boldsymbol{x}_2^{(2)}, ..., \boldsymbol{x}_2^{(n_1)}]^\top \in \mathbb{R}^{n_1 \times d_2}$. We learn a representation from function class $\mathcal{H}_1^{d_2} := \{f : \mathcal{X}_1 \to \mathbb{R}^{d_2}, [f(\cdot)] \in \mathcal{H}_1, \forall i \in [d_2]\}$ by using n_1 samples:

$$\tilde{\psi} := \text{argmin}_{f \in \mathcal{H}_1^{d_2}} [\frac{1}{n} \| \boldsymbol{X}_2 - f(\boldsymbol{X}_1^{pre}) \|_F^2].$$

For downstream tasks we similarly define $\boldsymbol{X}_1^{down} \in \mathbb{R}^{n_2 \times d_1}$, $Y \in \mathbb{R}^{n_2 \times d_3}$, and learn a linear classifier trained on $\tilde{\psi}(\boldsymbol{X}_1^{down})$:

$$\hat{W} = \text{argmin}_{\boldsymbol{W}} \frac{1}{2n_2} \| Y - \tilde{\psi}(\boldsymbol{X}_1^{down}) \boldsymbol{W} \|_F^2,$$

and

$$ER_{\tilde{\psi}}(\hat{\boldsymbol{W}}) := \mathbb{E}_{X_1} \| f_{\mathcal{H}_1}^\star - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1) \|_2^2$$

where $f_{\mathcal{H}_1}^\star = \mathbb{E}^L[Y | \phi_1(X_1)]$ is the best prediction inside $\mathcal{H}_1^{d_3}$.

Finite samples for both pretext and downstream tasks.

Assumption

Suppose there exists latent variable $Z \in \mathcal{Z}$ that ensures $\|\mathcal{C}_{\phi_1\phi_1}^{-1/2}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{Y}}}\|_{HS} \leq \epsilon$, and $\mathcal{C}_{\phi_{\bar{Y}}X_2}$ is full column rank, $\|\mathcal{C}_{Y\phi_{\bar{Y}}}\mathcal{C}_{\phi_{\bar{Y}}X_2}^{\dagger}\| = 1/\beta$. where $A^{\dagger}$ is pseudo-inverse, and $\phi_{\bar{y}}$ is the one-hot embedding for $\bar{Y} = [Y, Z]$.

The residual term $N := Y - \mathbb{E}[Y|X_1]$ is mean zero and assumed to be σ^2-subgaussian. When we use non-universal features ϕ_1, $\mathbb{E}[Y - f^{\star}(X_1)|X_1]$ might not be mean zero. We thus additionally assume a bounded $a := f^{\star} - f_{\mathcal{H}_1}^{\star} = \mathbb{E}[Y|X_1] - \mathbb{E}^{L}[Y|\phi_1(X_1)]$.

Assumption

There exists a universal constant b , such that

$$\|\mathcal{C}_{\phi_1\phi_1}^{-1/2}\phi_1(X_1)a(X_1)\|_F \leq b\sqrt{k}.$$

Main Result

Theorem

For a fixed $\delta \in (0,1)$, under the above Assumption, if $n1, n2 >> \rho^4(d_2 + log1/\delta)$, and we learn the pretext tasks such that: $\mathbb{E}\|\tilde{\psi}(X_1) - \psi^\star(X_1)\|_F^2 \leq \epsilon_{pre}^2$. Then we achieve generalization for downstream task with probability $1-\delta$:

$$ER_{\psi^\star}(\hat{\boldsymbol{W}}) \leq \mathcal{O}(\sigma^2 \frac{(d_2 + \log(d_2/\delta))}{n_2}) + \frac{\epsilon^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2}).$$

Remark

Our learned representation $\tilde{\psi} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ captures the information for Y with cardinality $k < d_2$. Therefore we could simply select the most important features to predict Y . Specifically, if we do PCA on $\tilde{\psi}(\boldsymbol{X}_1^{down})$ and use the top k features to predict Y , we could further improve the bound to

$$ER_{\psi^\star}(\hat{\boldsymbol{W}}) \leq \mathcal{O}(\sigma^2 \frac{(k + \log(k/\delta))}{n_2}) + \frac{\epsilon^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2}).$$

proof

Let $\Psi^\star$, L, E, V be defined as follows: Let $V = f^\star(\boldsymbol{X}_1^{down}) \equiv f^\star_{\mathcal{H}_1}(\boldsymbol{X}_1^{down}) \equiv \phi(\boldsymbol{X}_1^{down})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 Y}$. Denote

$$\begin{aligned}\Psi^\star &:= \psi^\star(\boldsymbol{X}_1^{down}) = \phi((\boldsymbol{X}_1^{down}))\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2} \\ &= \phi(\boldsymbol{X}_1^{down})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1\phi_{\bar{y}}}\mathcal{C}_{\phi_{\bar{y}}X_2} + \phi(\boldsymbol{X}_1^{down})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}} \\ &:= L + E.\end{aligned}$$

In this proof, we denote S_Y as the matrix such that $S_Y\phi_{\bar{y}} = Y$. Specifically, if Y is of dimension d_3, S_Y is of size $d_3 \times |\mathcal{Y}||\mathcal{Z}|$. Therefore $S_Y\Sigma_{\phi_y A} = \Sigma_{YA}$ for any random variable A. Therefore, similarly we have:

$$L\Sigma^\dagger_{X_2\phi_{\bar{y}}}\Sigma_{\phi_{\bar{y}}\phi_{\bar{y}}}S_Y^\top = L\Sigma^\dagger_{X_2\phi_{\bar{y}}}\Sigma_{\phi_{\bar{y}}Y} = L\bar{W} = V$$

where $\bar{\boldsymbol{W}} := \Sigma^\dagger_{X_2\phi_{\bar{y}}}\Sigma_{\phi_{\bar{y}}Y}$ satisfies $\|\bar{\boldsymbol{W}}\|_2 = 1/\beta$.

proof

We note that $E = \phi(\boldsymbol{X}_1^{down})\mathcal{C}_{\phi_1\phi_1}^{-1}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}}$ concentrates to $\mathcal{C}_{\phi_1\phi_1}^{-1/2}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}}$.
Specifically, when $n >> c + \log 1/\delta$, $\|E\|_F = \mathcal{O}(\epsilon)$.

In addition, write $\boldsymbol{E}^{pre} = \Psi - \Psi^\star = \tilde{\psi}(\boldsymbol{X}_1^{down}) - \psi^\star(\boldsymbol{X}_1^{down})$.
$\mathbb{E}\|\tilde{\psi}(X_1) - \psi^\star(X_1)\|^2 \leq \epsilon_{pre}$ impies that $\frac{1}{\sqrt{n_2}}\|\boldsymbol{E}^{pre}\| = \mathcal{O}(\epsilon_{pre})$.

proof

Note that,

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{N} + f^{\star}(\boldsymbol{X}_1^{down}) \\ &= \boldsymbol{N} + \Psi^{\star}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{down}) \\ &= \boldsymbol{N} + (\Psi + \boldsymbol{E}^{pre})\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{down}) \\ &= \Psi\bar{\boldsymbol{W}} + (\boldsymbol{N} + \boldsymbol{E}^{pre}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{down})).\end{aligned}$$

As $\|\boldsymbol{Y} - \Psi\hat{\boldsymbol{W}}\|_F^2 \leq \|\boldsymbol{Y} - \Psi\bar{\boldsymbol{W}}\|_F^2$,

$$\frac{1}{2n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F^2 \leq \frac{1}{n_2}\langle\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \boldsymbol{N} + \boldsymbol{E}^{pre}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{down})\rangle,$$

We analysis there terms seperately.

proof

$$\langle \Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \boldsymbol{N} \rangle \leq \|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F \|\mathbb{P}_\Psi \boldsymbol{N}\|_F = \mathcal{O}(\sqrt{d_2} + \log(d_2/\delta)\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F).$$

$$\langle \Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \boldsymbol{E}^{pre}\bar{\boldsymbol{W}} \rangle = \mathcal{O}(\frac{\sqrt{n}\epsilon_{pre}}{\beta}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F).$$

Write Ψ as $\phi(\boldsymbol{X}_1^{down})\boldsymbol{B}$, we have:

$$\begin{aligned}\langle \Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), a(\boldsymbol{X}_1^{down}) \rangle &\leq \sqrt{n_2 d_2}\|\mathcal{C}_{\phi_1}^{-1/2}\boldsymbol{B}(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F \\ &= \mathcal{O}(\sqrt{d_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F).\end{aligned}$$

proof

As a result,

$$\frac{1}{2\sqrt{n}}\|\Psi\hat{\boldsymbol{W}} - f^{\star}_{\mathcal{H}_1}(\boldsymbol{X}_1^{down})\|_F + \|\boldsymbol{E}\bar{\boldsymbol{W}}\| = \mathcal{O}(\frac{d + \log d_2/\delta}{n_2} + \frac{\epsilon + \epsilon_{pre}}{\beta}).$$

Finally, by concentrating $\frac{1}{n_2}\Psi^\top\Psi$ to $\mathbb{E}[\tilde{\psi}(X_1)\tilde{\psi}(X_1)^\top]$ we derive the desired result.

Outline

Simulations

Data generation.

- $d_1 = 50,\ d_2 = 40$;
- $\mu_{10}, \mu_{11} \in \mathbb{R}^{d_1} \sim U([0,1]^{d_1})$ and $\mu_{20}, \mu_{11} \in \mathbb{R}^{d_2} \sim U([0,1]^{d_2})$;
- $Y \sim U\{0,1\}$.
- $X_1 \sim (1-Y)\mathcal{N}(\mu_{10}, \boldsymbol{I}) + Y\mathcal{N}(\mu_{11}, \boldsymbol{I})$;
- $X_2 \sim (1-Y)\mathcal{N}(\mu_{20}, \boldsymbol{I}) + Y\mathcal{N}(\mu_{21}, \boldsymbol{I})$.

We sample the pretext dataset $\{\boldsymbol{x}_1^{(i,pre)}, \boldsymbol{x}_2^{(i)}\}_{i=1}^{n_1}$ to learn ψ and sample $\{\boldsymbol{x}_1^{(i,down)}, y\}_{i=1}^{n_2}$ to learn $\boldsymbol{W}$, the linear classifier for downstream task of predicting Y .

Simulations

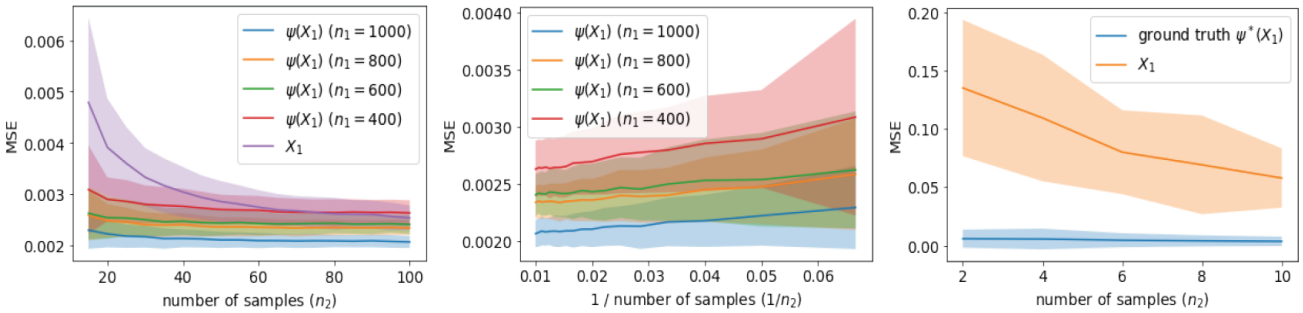

Figure 1: **Left**: MSE of using ψ to predict Y versus using X_1 directly to predict Y. Using ψ consistently outperforms using X_1. **Middle**: MSE of ψ learned with different n_1. The MSE scale with $1/n_2$ as indicated by our analysis. **Right**: MSE of using the optimal ψ^* (i.e., $\mathbb{E}[X_2|X_2]$) to predict Y versus using X_1 directly to predict Y. The ground truth ψ^* gets almost zero MSE with very few samples. Simulations are repeated 100 times, with the mean shown in a solid line and one standard deviation shown in the shadow.

Computer Vision Task

We test if learning with ψ is more effective than learning directly with X_1, in a realistic setting (without enforcing conditional independence). Specifically, we test on the Yearbook dataset, where inputs are pictures of people from yearbooks and goal is to predict the year when the pictures are taken (denoted as Y), which ranges from 1905 to 2013. We resize all the portal to be 128 by 128. We crop out the center 64 by 64 pixels (the face), and treat it as X_2, and treat the outer rim as X_1 as shown in Figure 3 on the left. Our task is to predict Y , which is the year when the portraits are taken.

Computer Vision Task

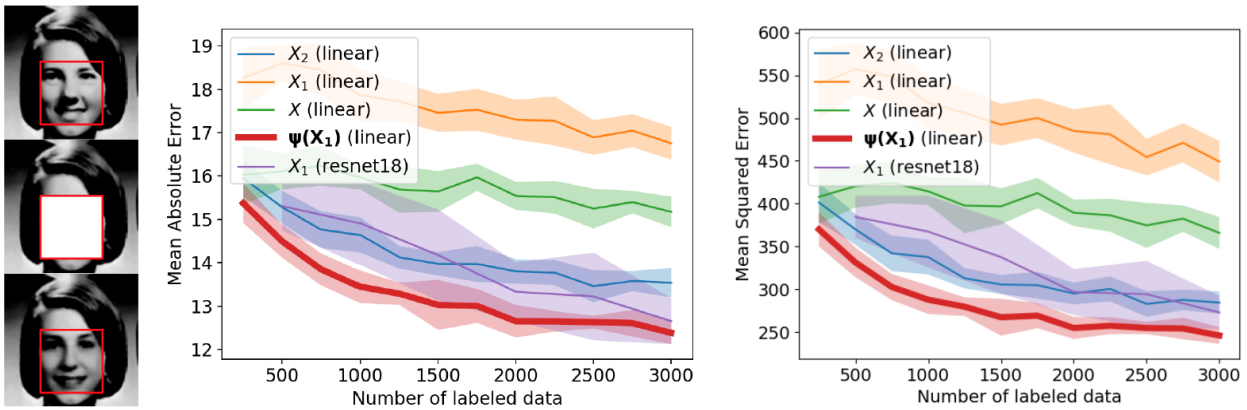

Figure 3: **Left**: Example of the X_2 (in the red box of the 1st row), the X_1 (out of the red box of the 1st row), the input to the inpainting task (the second row), $\psi(X_1)$ (the 3 row in the red box), and in this example $Y = 1967$. **Middle**: Mean Squared Error comparison of yearbook regression. **Right**: Mean Absolute Error comparison of yearbook regression. Simulations are repeated 10 times, with the mean shown in solid line and one standard deviation shown in shadow.

Future Directions

- Allow (feature map $\rightarrow$ representation) to be non-linear;
- to derive a lower bound of $ER(\hat{W})$;
- Consider to use neural network in pretext tasks.

THANK YOU!